Discriminant Interest Points are Stable

Dashan Gao Nuno Vasconcelos
Department of Electrical and Computer Engineering
University of California, San Diego
dgao@ucsd.edu, nuno@ece.ucsd.edu

Abstract

A study of the performance of recently introduced discriminant methods for interest point detection [6, 14] is presented. It has been previously shown that the resulting interest points are more informative for object recognition than those produced by the detectors currently used in computer vision. Little is, however, known about the properties of discriminant points with respect to the metrics, such as repeatability, that have been traditionally used to evaluate interest point detection. A thorough experimental evaluation of the stability of discriminant points is presented, and this stability compared to those of four popular methods. In particular, we consider image correspondence under geometric and photometric transformations, and extend the experimental protocol proposed by Mikolajczyk et al. [13] for the evaluation of stability with respect to such transformations. The extended protocol is suitable for the evaluation of both bottom-up and top-down (learned) detectors. It is shown that the stability of discriminant interest points is comparable, and frequently superior, to those of interest points produced by various currently popular techniques.

1. Introduction

Saliency mechanisms play an important role in the ability of biological vision systems to perform complex tasks, such as detection and recognition of objects from cluttered background. By identifying certain regions of the visual field as more important, or *salient*, than others they enable a non-uniform allocation of perceptual resources that eases the computational burden posed by visual tasks [15]. In the computer vision literature, the saliency problem is frequently referred to as the extraction of *interest points*, and has been a subject of research for a few decades. Various interest point detectors have been proposed and applied to many computer vision tasks. Recently, interest point detectors have been extensively used in the context of the extraction of image descriptors for matching-based recognition [9, 10], or for learning object categories [4, 1].

Although it has long been known that human judgements of saliency [18] can be of two types, *bottom-up* (stimulus-driven) or *top-down* (task-specific), most popular interest point detectors have an exclusively bottom-up nature. These detectors define interest points as image locations that exhibit some properties which are universally desirable [7, 16, 8, 12, 3]. The most popular among such properties is that of stability to various imaging and geometrical transformations, an optimality criterion for which many interest point detectors have been proposed [7, 5, 17, 12, 10]. Recently, it has been shown that some of these detectors are stable with respect to a broad class of image transformations, including various geometric transformations (rotation, scaling, affine mappings), lighting variation, blurring, and image compression [13]. When compared with the top-down strategies, bottom-up detection has various advantages, including 1) optimality criteria that are amenable to closed-form mathematical solutions, 2) freedom from computationally intensive training, and 3) low implementation complexity. These advantages are, however, closely tied to what is also their major limitation: due to the absence of task-specific focus, bottom-up detectors *can only* be optimal in very generic senses, and the resulting interest points are usually not the best for any specific application.

Recently, there have been attempts to overcome this problem, by introducing top-down interest point detectors [6, 14]. A proposal which explicitly targets recognition is to define saliency in terms of features. The basic idea is that, for recognition, salient features are those which best discriminate between the class to recognize and the remaining classes [6]. Interest points are then the locations where those features have maximal response. This tunes the interest point detector to the class to recognize, guaranteeing that only image locations which are *informative for the recognition of that class* are denoted of interest. Experimental evaluation, in the domain of learning object categories from cluttered imagery, has shown that these top-down interest points can be significantly more informative for recognition than those produced by bottom-up methods [6, 2].

It is, however, not clear how the two approaches compare

1-4244-1180-7/07/$25.00 ©2007 IEEE

with respect to the criteria, such as stability, traditionally used to evaluate interest points. One potential limitation of the top-down strategy lies in its dependence on a training set. For applications where training images are scarce, top-down detectors could have poor generalization, in which case the resulting interest points would likely be very unstable. Unlike bottom-up detectors, whose stability has been thoroughly characterized, the robustness of top-down interest point detectors has not yet been investigated in detail.

In this work, we present the results of a study of this question, through a detailed experimental evaluation of the repeatability of discriminant interest points. In particular, we consider image correspondence under various geometric and photometric transformations. The main contributions are as follows. First, we extend the experimental protocol proposed in [13] for the evaluation of interest point detection, so as to make it applicable to both bottom-up and top-down (learning-based) detectors. The extended protocol consists of a series of experiments which quantify the relationship between stability and training set size. The original protocol becomes a special case of this extension, enabling a fair comparison of bottom-up and top-down techniques. The second contribution is a comparison between the stability of discriminant interest points and those of interest points produced by four bottom-up detectors widely used in computer vision: the scale saliency detector (SSD) of [8], the Harris-Laplace (HarrLap) detector of [12], the Hessian-Laplace (HesLap) detector of [12], and the maximally stable extremal region (MSER) detector of [10].

It is shown that the matching performance of the top-down interest points is comparable with, and frequently superior to, those of their bottom-up counterparts. In particular, discriminant points perform best when trained with rich training sets, and achieve comparable performance when training sets are small. Only in the extreme cases of complex scenes, subject to substantial transformations, and very little training data, did their stability scores dropped below those of the best non-discriminant methods. This suggests that the discriminant formulation is a good idea even from a stability point of view: not only it enables the design of detectors that can be made more invariant by simply increasing the richness of their training sets, but it appears to work well even when training data is limited. We propose an explanation for this observation, based on connections between stability, sparseness, and discrimination.

2. Interest point detectors

We start with a brief review of all interest point detectors compared in the experiments of the following sections[1].

2.1. Discriminant saliency detector (DSD)

Under the discriminant formulation [6], interest points are the locations of maximal response of a set of salient visual features. Feature saliency is defined as the discriminant power of a feature with respect to the classification problem that opposes the class of interest to all other classes in the recognition problem. In the implementation of [6], the set of candidate features consists of the coefficients of the discrete cosine transform (DCT) and salient features are selected by the maximum marginal diversity (MMD) criterion. In particular, images are projected into a K-dimensional feature space, and the marginal distribution of each feature (X_k) under each class (Y), $P_{X_k|Y}(x|i), i \in \{0,1\}, k \in \{0,\ldots,K-1\}$, is estimated. Features are then sorted by decreasing mutual information with the class label,

$$I(X_k;Y) =< KL[P_{X_k|Y}(x|i)||P_{X_k}(x)] >_Y,$$

where $< f(i) >_Y = \sum_{i=1}^{M} P_Y(i)f(i)$, and $KL[p||q] = \int p(s)\log\frac{p(x)}{q(x)}dx$ the Kullback-Leibler divergence between p and q. The features of largest mutual information are finally selected. A saliency map is generated by 1) projecting the image into the subspace spanned by the salient features, and 2) combining the resulting projections $R_i(\mathbf{x})$ according to

$$S_D(\mathbf{x}) = \sum_{i=1}^{n} \omega_i R_i^2(\mathbf{x}),$$

with ω_i set to the mutual information between the corresponding features and class label. Interest points are then determined by a non-maximum suppression stage, which sets the scale of each point to the spatial support of the feature of largest response at that point. Figure 1 (b) illustrates some salient locations generated by DSD.

2.2. Scale saliency detector (SSD)

This detector defines saliency as spatial unpredictability of image regions, and relies on the changes of the information content of the distribution of image intensities over spatial scales to detect interest point locations [8]. To detect an interest point, the entropy, $H(s,\mathbf{x})$, of the histogram of local intensities over the image neighborhood of circular scale s, centered at $\mathbf{x}$, is first computed,

$$H(s,\mathbf{x}) = -\sum_{I} p(I,s,\mathbf{x})\log p(I,s,\mathbf{x}),$$

where $p(I,s,\mathbf{x})$ is the histogram of image intensities. Its local maximum over scales, $H(\mathbf{x})$, is then determined and the associated scale is considered as a candidate scale, s_p,

[1] Implementations of all detectors are available from their original authors, and were used to produce all results presented in this paper. Binary codes are available from *http://www.svcl.ucsd.edu/projects/saliency*, *http://www.robots.ox.ac.uk/~timork/salscale.html* and *http://www.robots.ox.ac.uk/~vgg/research/affine/detectors.html*. The default parameters settings, if applicable, were used.

for location $\mathbf{x}$. A saliency map is obtained as a weighted entropy,

$$S_S(\mathbf{x}) = H(\mathbf{x})W(s_p, \mathbf{x})$$

where $W(s, \mathbf{x}) = s \int \left| \frac{\partial}{\partial s} p(I, s, \mathbf{x}) \right| \mathrm{d}I$, and interest points are finally located by clustering this saliency map. An example of salient points detected by the SSD is presented in Figure 1 (c).

2.3. Harris-Laplace (HarrLap) and Hessian-Laplace (HesLap) detectors

These two detectors are designed to maximize the stability of image regions to some geometric transformations (e.g. scaling) [12]. They are essentially corner detectors, inspired by the observation that corners are stable under various types of image transformations. In particular, to find interest points, the HarrLap detector relies on an auto-correlation matrix [7],

$$H_a(x, y) = \sum_{(u,v)} w_{u,v} \nabla I(x+u, y+v) \nabla^T I(x+u, y+v) \tag{1}$$

where $\nabla I(\mathbf{x}) = (I_x(\mathbf{x}), I_y(\mathbf{x}))^T$, is the spatial gradient of the image at location $\mathbf{x} = (x, y)$, and $w_{u,v}$ a low-pass filter (typically a Gaussian) that smoothes the image derivatives. HesLap, on the other hand, uses the Hessian matrix of local image intensities,

$$H_e(x, y) = \sum_{(u,v)} w_{u,v} \begin{bmatrix} I_{xx}(x+u, y+v) & I_{xy}(x+u, y+v) \\ I_{yx}(x+u, y+v) & I_{yy}(x+u, y+v) \end{bmatrix}, \tag{2}$$

where $I_{xy}(\cdot)$ represents the second order derivative with respect to x and y. In both cases, the scale of interest points is determined by application of the Laplacian operator,

$$|L(x, y, s)| = s^2 |I_{xx}(x, y, s) + I_{yy}(x, y, s)|,$$

for a number of scales s. This operator has been demonstrated to give the best scale selection results in the experimental comparison of [11]. Starting from a set of initial locations obtained by detecting local maxima of (1) or (2) on multiple scales, the final interest points are detected by an iterative algorithm, which sequentially searches for local maxima over scale and space [12]. Examples of interest points detected by HarrLap and HesLap are shown in Figure 1 (d) and (e).

2.4. Maximally stable extremal region (MSER) detector

The maximally stable extremal region (MSER) detector defines interest regions by an *extremal* property of image intensity. The word *extremal* refers to the property that all pixels inside the MSER have either higher (brighter) or lower (darker) intensity than all the pixels in its outer boundary. The set of extremal regions is the set of all connected components obtained by thresholding a gray-level image, and possesses some desirable properties: the set is closed under 1) any continuous (and projective) transformation of image coordinates, or 2) any monotonic transformation of image intensities. The "maximally stable" extremal regions are the ones whose area changes the least in response to a change of threshold.

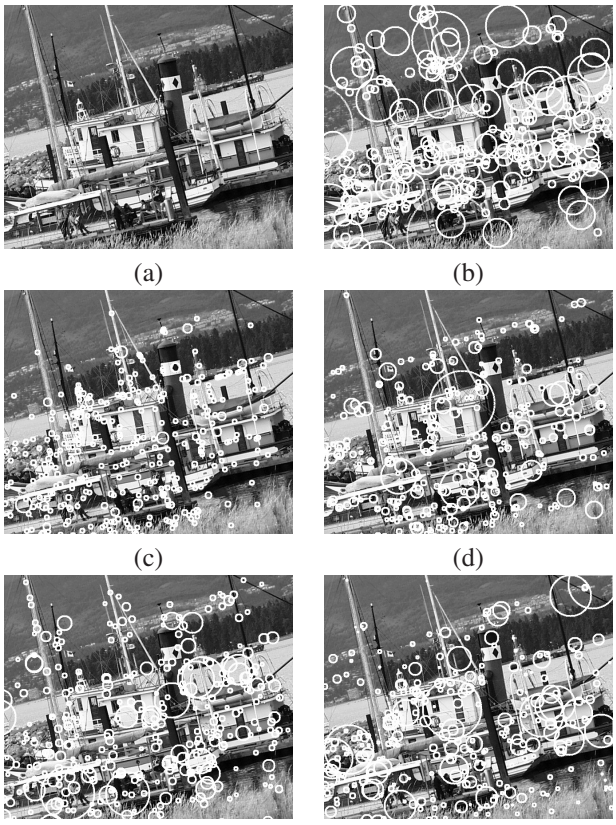

Figure 1. Interest points detected, on image (a), by different approaches: (b) DSD, (c) SSD, (d) HarrLap, (e) HesLap, (f) MSER. For intelligibility, the points shown are randomly selected from all locations generated by the detectors.

The enumeration of extremal regions can be implemented very efficiently, with complexity $O(n \log \log n)$, where n is the number of image pixels [10]. The implementation of the complete MSER detector is as follows. First, pixels are sorted by intensity, and sequentially placed in the image (either in descending or ascending order). The list of connected components, and their area, is stored as a function of pixel intensity. The intensity levels which are local minima of the rate of area change are finally selected as thresholds, and used to produce the MSERs. Even though the output regions produced by MSER can be of arbitrary shape, they (like those of other detectors) are represented

by a circular shape, for comparison purposes. The regions produced by MSER are illustrated in Figure 1 (f).

3. Repeatability

The performance of discriminant saliency, in the context of object detection from cluttered scenes, has been studied in [6]. This study has shown that discriminant saliency produces interest points which are more informative about the location of the objects to recognize than all other interest point detectors. This is not completely surprising, given that discriminant saliency detectors are optimized for the class of interest. In what follows we evaluate performance with respect to the task for which the remaining detectors are optimal, or close to optimal. This is the stability of interest points with respect to various generic image transformations.

3.1. Experimental protocol

Ideally, interest points should be unaffected by changes of the various (scene-independent) parameters which control the imaging process, e.g. lighting, geometric transformations (such as rotation and scaling), and so forth. Mikolajczyk et al. [13] have devised an experimental protocol for evaluating the repeatability of interest points under various such transformations. The protocol includes 8 classes of transformations, each class consisting of 6 images produced by applying a set of transformations, from the same family, to a common scene. The transformations include joint scaling and rotation, changes of viewpoint angle (homographies), blurring, JPEG artifacts, and lighting. All detectors other than DSD are explicitly designed to be invariant to, at least, some of these variations. Scale + rotation, view point changes, and blurring are applied to sets of two scenes, which can be roughly characterized as textured (e.g. images of tree bark or of a brick wall) or structured (e.g. an outdoors scene depicting a boat or a wall covered with graffiti).

3.2. Extending the protocol for learning

Since the protocol of [13] does not define training and test images, we propose an extension applicable to learning-based methods, such as discriminant saliency. This extended protocol is based on various rounds of experiments. At the k^{th} round, the first k images of a given class are treated as a training set for that class, and the repeatability scores of the learned interest point detector are measured on the remaining $6 - k$ images. This is accomplished by matching the interest points detected on these images to the reference image, which is the k^{th} image. When $k = 1$, this reduces to the protocol of [13], but larger values of k enable a quantification of the improvement of stability with the richness of the training set. The new protocol is illustrated in Figure 2 for $k = 1$ and 2. In the experiment reported below, the repeatability score of DSD was measured for $k = \{1, 2, 3\}$, and compared to the other detectors operating under the same test protocol (i.e., using image k as a reference). In preliminary experiments, we noted that the implementation of DSD of [6] could not deal with the large variations of scale of this dataset. To overcome this, we implemented a simple multi-resolution extension of DSD: discriminant interest points were first detected at each layer of a Gaussian pyramid decomposition of the image and, at each interest point location, the layer of largest saliency was selected. This type of processing is already included in all other detectors.

3.3. Repeatability score

The criterion used to find the corresponding points between a pair of images is that of [13]. In particular, a mapped interest point, R_a, is considered to match the reference image if there exists an interest point, R_f, in the latter for which the *overlap error* is sufficiently small, i.e.

$$1 - \frac{R_a \cap R_f}{R_a \cup R_f} < \epsilon, \tag{3}$$

where $\cap$ represents intersection, and $\cup$ union. To avoid favoring matches between larger interest points, the reference region was normalized to a radius of 30 pixels before matching, as suggested by [13]. The matching threshold, ϵ, was set to 0.4. The repeatability score for a given pair of images is computed, as in [13], as the ratio between the number of correspondences and the smaller of the number of regions in the pair.

3.4. Results

The average repeatability scores obtained (across the set of test images) by each interest point detector are shown, as a function of the reference image number k, in Figure 3.

The most surprising conclusion that can be taken from the figures is the good performance of DSD. When the training set includes multiple positive examples ($k > 1$), it outperforms all other methods for 7 of the 8 classes. But even when the set of positive training examples is a single image (i.e., there are no training examples for the deformations suffered by each image patch under the transformations considered) DSD is competitive with all other techniques. In fact, it still achieves the top repeatability scores for five of the eight classes (Figure 3 (d)-(h)), and is very close to the best for another (Figure 3 (b)). HesLap achieved the best performance among the non-discriminant interest point detectors.

It is also interesting to compare the relative performance of DSD and HesLap by transformation and image class. With respect to transformations, DSD is the most robust

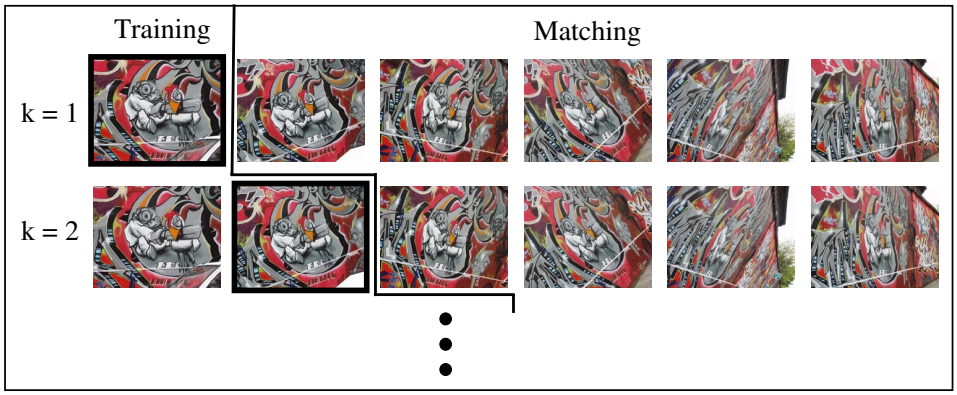

Figure 2. Extended protocol for the evaluation of the repeatability of learned interest points. At the k^{th} round, the detector is trained on the first k images, and the repeatability score measured by matching the remaining images to the reference, which is set to the last training image, and shown with thick boundaries.

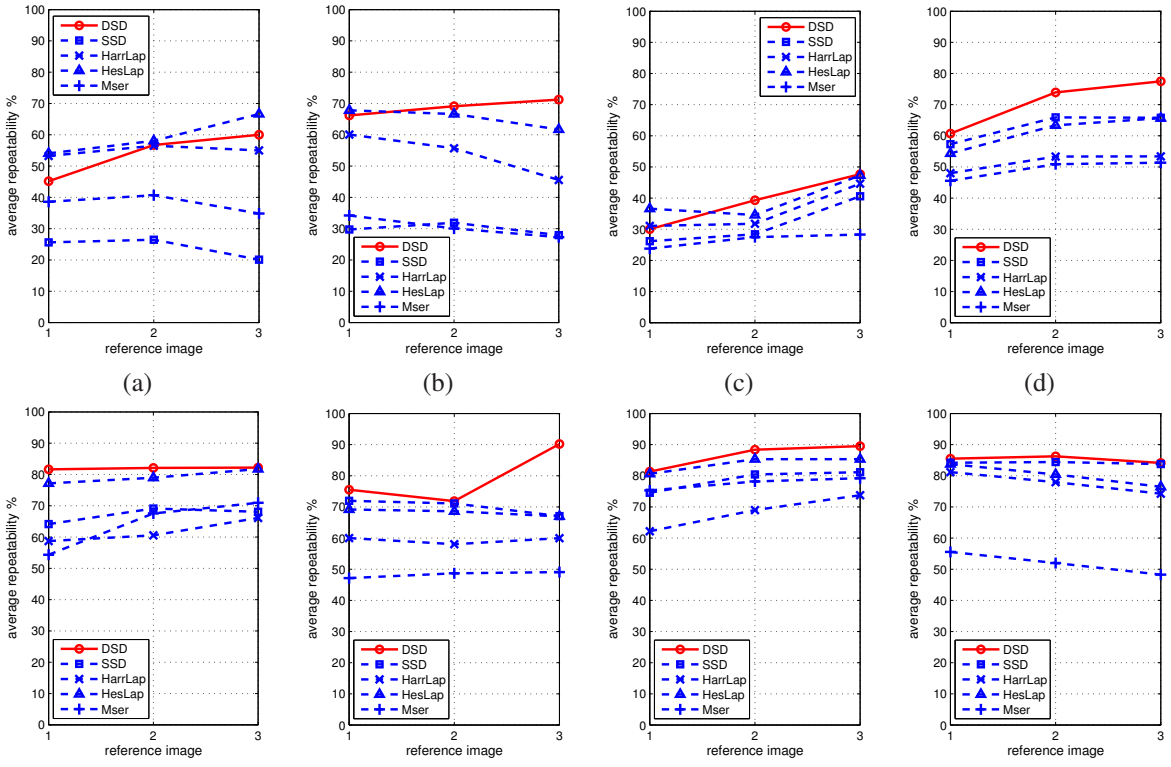

Figure 3. Repeatability of interest points under different conditions: *scale + rotation* ((a) for structure & (b) for texture); *viewpoint angle* ((c) for structure & (d) for texture); *blur* ((e) for structure & (f) for texture); *JPEG compression* (g); and *lighting* (h).

method in the presence of blurring, JPEG artifacts and lighting transformations (Figure 3 (e-h)) independently of the degree of training. It also achieves the best performance for changes of viewpoint angle, but this can require more than one positive example (Figure 3 (c)). Its worst performance occurs under combinations of scale and rotation (Figure 3 (a) & (b)), where it is always inferior to that of HesLap for small amounts of training, and sometimes inferior for the largest training sets. With respect to image class, the robustness of DSD to geometric transformations is better for texture (Figure 3 (b) & (d)) than for structured scenes (Figure 3 (a) & (c)). While, for the former, DSD achieves the best, or close to the best, performance at all training levels, for structured scenes DSD is less invariant than HesLap for the majority of the training regimes.

4. Discussion

Overall, the results above illustrate some of the trade-offs associated with learning-based (top-down) interest point detectors, such as DSD. On one hand, the ability to select specific features for the class under consideration, increases not only the discriminant power but also the stability of interest point detection. It appears that the principle of discriminant learning is a good idea even from a repeatability point of view. It enables the design of detectors which can be made more invariant by simply increasing the richness of the transformations covered by their training sets. This is a property that bottom-up routines lack, and sometimes leads to dramatic variability of repeatability scores across classes (see the curves of SSD on Figure 3 for an example). On the other hand, the generalization of a top-down detector depends on the quality of the training data and the complexity of the mappings that must be learned. In Figure 3, this can be seen by the consistent performance loss associated with smaller training sets, and the greater difficulties posed by structured scenes, when compared to texture. When little training data is available, or the mappings have great complexity, explicit encoding of certain types of invariance (as done by the bottom-up detectors) can be more effective. In this sense, the combination of top-down and bottom-up interest point detectors, to optimally balance the trade-off between learning and pre-specification of invariance, could be beneficial.

It is, nevertheless, surprising that, with respect to stability, DSD is competitive with the existing interest point detectors, even for quite small amounts of training data. In the experiments above, competitive performance was achieved for $k = 1$, i.e. the case where only one image is available for training. In this case, the training set contains no examples of the variability that a pattern may present in response to the transformations considered. The ability of the discriminant detector to learn stable points under these circumstances is probably due to a strong connection between the stability of a pattern and its frequency of occurrence in the universe of natural images.

Stability requires 2D patterns of image intensity (e.g. corners), which are less frequent than one dimensional patterns (e.g. lines), and even less frequent than flat patterns (flat surfaces). Hence, stable features are more rare than non-stable features, making their presence in an object category more discriminant. This makes them more likely to be selected than non-stable features, even when the training set contains no explicit information about the deformations that they undergo, as the image is subject to transformations. This is illustrated in Figure 4, where we present a number of discriminant interest points extracted from the wall image, for $k = 1$. Note that the interest points tend to be co-located with corners and T-junctions, which are both rare (in natural images) and stable.

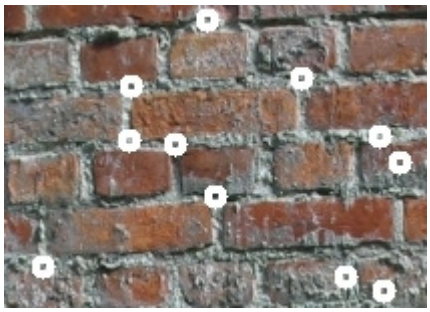

Figure 4. Discriminant interest points learned from a single training image. The points are co-located with image patterns that are both rare and stable.

Acknowledgments

This research was supported by NSF award IIS-0448609.

References

[1] S. Agarwal and D. Roth. Learning a sparse representation for objection detection. In *Proc. ECCV*, volume 4, pages 113–130, 2002.

[2] A. Bar-Hillel, T. Hertz, and D. Weinshall. Efficient learning of relational object class models. In *IEEE Intl. Conf. on Computer Vision*, volume 2, pages 1762–1769, 2005.

[3] N. D. Bruce and J. K. Tsotsos. Saliency based on information maximization. In *Proc. NIPS*, 2005.

[4] R. Fergus, P. Perona, and A. Zisserman. Object class recognition by unsupervised scale-invariant learning. In *Proc. IEEE Conf. CVPR*, 2003.

[5] W. Förstner. A framework for low level feature ex-traction. In *Proc. of ECCV*, pages 383–394, 1994.

[6] D. Gao and N. Vasconcelos. Discriminant saliency for visual recognition from cluttered scenes. In *Proc. NIPS*, pages 481–488, 2004.

[7] C. Harris and M. Stephens. A combined corner and edge detector. In *Alvey Vision Conference*, pages 147–151, 1988.

[8] T. Kadir and M. Brady. Scale, saliency and image description. *Int'l. J. Comp. Vis.*, 45:83–105, Nov. 2001.

[9] D. G. Lowe. Object recognition from local scale-invariant features. In *Proc. ICCV*, pages 1150–1157, 1999.

[10] J. Matas, O. Chum, M. Urban, and T. Pajdla. Robust wide-baseline stereo from maximally stable extremal regions. *Image and Vision Computing*, 22(10):761–767, September 2004.

[11] K. Mikolajczyk and C. Schmid. Indexing based on scale invariant interest points. In *Proc. ICCV*, pages 525–531, 2001.

[12] K. Mikolajczyk and C. Schmid. Scale and affine invariant interest point detectors. *Int'l J. Comp. Vis.*, 60(1):63–86, 2004.

[13] K. Mikolajczyk, T. Tuytelaars, C. Schmid, A. Zisserman, J. Matas, F. Schaffalitzky, T. Kadir, and L. V. Gool. A comparison of affine region detectors. *Int'l J. Comp. Vis.*, 65:43–72, 2005.

[14] V. Navalpakkam and L. Itti. Optimal cue selection strategy. In *Proc. NIPS*, pages 987–994, 2005.

[15] S. E. Palmer. *Vision Science: Photons to Phenomenology*. The MIT Press, 1999.

[16] A. Sha'ashua and S. Ullman. Structural saliency: the detection of globally salient structures using a locally connected network. In *Proc. ICCV*, pages 321–327, 1988.

[17] J. Shi and C. Tomasi. Good features to track. In *Proc. IEEE Conf. CVPR*, pages 593–600, 1994.

[18] A. Yarbus. *Eye movements and vision*. Plenum, New York, 1967.